The Preservation and Use of Machine Readable Records

Paper for the
Conference on the National Archives and Statistical Research
May 27 and 28, 1968

by Richard Ruggles
Yale University

The role of archives in economic and social research

In the study of economics, I have always been fascinated by the fact that although economists are primarily interested in the future course of events they do recognize that such analysis is purely speculative, and is based on fragmentary and uncertain knowledge of the present. Furthermore, by the time more reliable and relevant information can be obtained and analyzed, the present is past. Unfortunately there is a distressing tendency among many economists to consider information about the past as old wives' tales - or at best "old hat"; they are convinced that the present is radically different from the past, on the basis of little or no evidence. In spite of this, however, an understanding of the present and an ability to forecast the future must of necessity be based upon an understanding of the past and of the process of change. Unlike the physical scientist, the social scientist cannot set up laboratories or develop elaborate experiments to generate the basic information which he uses. He must make such observations as he can at points in time, and realize that they are dimensioned in time. Statistical data are the social scientist's primary tool of analysis, and in analyzing the process of change past data are quite as important as current data. By focusing attention on historical statistical resources, the National Archives and the National Academy of Sciences are highlighting the importance of this basic tool for the development of the social sciences.

The implication of the computer for the preservation and use of statistical data

The specific aspect of the problem which I would like to discuss this morning is the way in which the existence of the computer has affected the preservation and use of statistical data. I am going to refrain from telling you that there has been a computer revolution, but I understand the possibility has been rumored, and if one should develop it might have serious repercussions.

In the period B. C. (before computer), the National Archives may well have considered that it was flooded with information. In terms of pieces of paper, this is undoubtedly so. But in terms of statistical information in the social sciences, the body of useful archival material which was not already in published form was limited and fragmentary. It is true that the original data which underlay the published and unpublished statistics could have been retained, but there are practical and economic limitations to this. Thus for example it seems wasteful to retain warehouse after warehouse full of tax returns, or voluminous hand files of government agencies, when the cost of accessibility to such documents for statistical purposes is so high that no self-respecting social scientist could afford to indulge in analyzing these materials unless someone unwisely provided funds and he himself was completely devoid of the puritan ethic. For this reason the collection of data for the Archives understandably has given higher priority to the processed, aggregated, and tabulated statistical information which different government agencies have provided as summaries of their basic records. It is these records which in the past have provided the most valuable sets of data available to the social scientist. Even where the social scientist had large amounts of micro-data culled out for him, inability to process, retabulate, and analyze it made the value of its preservation highly questionable.

The introduction of electronic data processing has of course drastically changed this situation. For the social sciences, the original basic information in some edited, machine readable form provides a much better basis for research than do the summary aggregated tabulations. Aggregation by its very nature obscures the interrelationships in the data both for changes over time and among variables. Information on these interrelationships is required for testing hypotheses about economic and social behavior or changes in the structure of economic and social characteristics. Unfortunately, the value of micro-data sets has not been fully appreciated by all government agencies. Several years ago, when the Committee on the Preservation and Use of Data of the Social Science Research Council was examining this problem, they discovered that many government agencies preserved computer tapes of the aggregated tabulations but destroyed the tapes of their basic data. In some agencies this situation may have developed because those responsible for the data were ashamed of the quality of the basic data, and understandably they felt these should be destroyed before anyone else had the chance to see them.

The problem of what to save

Nevertheless, it is still true that there is a problem of what data should be saved. We do not wish to be buried in the debris of the information explosion. Not everything which is produced as a byproduct of the operation of the system is significant for research purposes. Even though I am an economist, let me hasten to add that I do not believe in assessing the value of information in terms of its market demand. On several occasions I have been somewhat dismayed when evaluation of sets of data is made to rest upon the question of how many people will use the information. The physical scientist, in conducting his experiments, does not use this market criterion in evaluating the usefulness of the information he generates. No one asks how many people use the flow of information that comes back from the moon. What is more important is what contribution such a flow of information can make to knowledge. If libraries took the position that they would acquire only those books which were in most demand, they would be filled up with textbooks and elementary books. In this connection, it is interesting to note that if you go through a university library and examine its collections in various areas you can fairly well build up a picture of the competence of the faculty during various periods in the past. For example, the Yale library reflects quite well the interests of

Irving Fisher, and at the same time it reveals significant gaps in major fields of economics during certain periods in the past. Unfortunately, it is difficult for librarians without substantive knowledge in various areas to build relevant and pertinent collections. As a substitute for knowledge of what is important, a librarian may sometimes employ the credo that quantity makes quality. In part this may be why our libraries are suffering from a shortage of space and require such sizeable budgets. It would be indeed unfortunate if the preservation of statistical data followed the same lines.

In other words, decisions on what data should be preserved should be made in terms of the value of the data for research purposes. This suggests in part that those making these decisions should themselves be professional research workers in the area, and that they should take the opportunity to consult widely with others who are also engaged in related specialized research.

However, I am somewhat disturbed by the lack of system and the appearance of chaos which results from such ad hoc decisions about data preservation. More research is required on the optimal design of information systems for economic and social research. At the present time, the highly decentralized federal statistical system produces a bewildering mass of obviously related but not always consistent information. There is a growing awareness that, at least with respect to information relating to economic transactions, a more systematic approach is needed.

The United Nations currently is proposing a revision in its standard system of national accounts which will embrace many other kinds of transactions data. Although one may raise questions about the specific way the United Nations is going about this, few if any national accountants dispute the desirability of developing a single, integrated system for economic data.

The integration of non-transaction data of a social, demographic, or institutional nature is less obvious. However, it is precisely in this dimension that the ability of the computer to process masses of micro-data gives us hope of ultimate solution. It is possible to attach to data about individual reporting units a great deal of information which is not necessarily dollars and cents. Thus for example, if one takes the household as a unit, information on composition of the household can be provided in terms of the age, sex, education, occupation, and employment of the members, together with related transaction information on income, expenditures, ownership of assets, existence of financial liabilities, etc. If the integrated economic accounting system is sectored along lines of reporting units such as households, enterprises, governmental units, etc., it will be possible to relate the micro-data sets to macro-economic data. For example, it is possible to construct a set of household sector accounts for the economy as a whole which summarize the income and expenditures of the households and their balance sheets. The set of micro-data underlying this account would consist of a sample of households which when properly weighted would yield the major economic constructs in the national accounts. The micro-data set would, however, have a great deal of additional information about each household and the members constituting it. Ideally it would be desirable to have panel micro-data so that information on the changes in individual households and their behavior over time would be inherent in the micro-data. This does require that the sectoring and classification systems used for the macro-economic system have direct correspondence to those of the micro-system. The value of such linkage is great. It would provide the social, demographic, and institutional information behind the economic data system. Economic analysis itself would be considerably enriched, since the behavior and structure of the system could be directly related to the social, institutional and demographic factors involved, and more realistic micro-models involving the technique of simulation could be developed for the testing of hypotheses and the analysis of policy.

The development of a general integrated statistical framework as a guide to the preservation of data has another use. It can not only help answer the question of what should be saved, but it can also point up serious gaps in our basic knowledge. For example, the Wealth Inventory Planning Study dramatically pointed up the gaps in knowledge in this important area. A comprehensive statistical framework, therefore, would help us obtain a systematic inventory of our knowledge.

The need for documentation

It is not enough, however, merely to preserve important data; the data must be preserved in a manner such that they can be used. I was very glad to see the emphasis which Mr. Alldredge has placed upon documentation. Machine readable information without the documentation required to interpret it is a tale told by an idiot. In the most extreme form the data becomes neither human readable nor machine readable.

The form in which most computer tapes now provide data is in large degree the result of the limitations of early computers, in terms of memory and capacity to deal with a wide variety of formats. When computer tapes were first used, they contained for the most part purely digital information packed as closely together as was feasible and without distinguishing characteristics. Early printouts often required the user to write in appropriate labels by hand, or to develop masks which could be laid on the printout to provide labels. As the computer has developed, there has been a vast improvement in the printouts produced, and considerable energy is now expended in providing programs which yield attractive, well laid out and fully labeled output.

Despite this improvement in the form and appearance of the final output, the characteristics of input tapes have not changed significantly. Users generally put on a computer tape simply a title as a header, followed by the date in as efficient a form as possible. In most cases there is also a label on the tape container giving an abbreviated title and a tape identification number. Information concerning the format, layout, arrangement of the tape, and the labels which should accompany it is sometimes written down in systematic form, but often the person using the tape must try to obtain this information from those responsible for erating it. As a result fragments of the necessary documentation are provided on a Xeroxed piece of paper which is easily lost or destroyed. It is common experience in using tapes obtained from others to find that insufficient information is available, and that a series of experiments is needed in order to find out how to use the tape. Sometimes these problems are so formidable that a research worker may consume most of the computer time allocated to his project simply in trying to find out how to use the tape. The basic problem is not so much one of failure of computer systems as it is the failure to provide the basic documentation which is required in useful form.

In order to overcome these difficulties, it would be useful if all essential documentation relating to a computer file were put into both human and machine readable form and placed at the beginning of the file. Such a documentation header would have three major functions. First, it would provide as part of the file itself the essential human readable documentation relating to the file. Reproduction of such basic human readable documentation would be automatic with any printout of the file.

There would be no problem of information getting lost or mismatched. Second, if every file had to be accompanied by well-defined human and machine readable documentation, those creating the file would be required to provide it. Much of the current problem is due to the fact that those creating machine readable files never record the necessary documentation. Third, developing machine readable documentation as part of machine readable data files will automatically provide the information necessary for processing the file. This will not only reduce error, but it will also make possible use of standard programs which can call on the machine readable documentation for the required information and thus allow the use of generalized rather than specialized programs.

The development of human and machine readable documentation can provide the same sort of discipline for computer files which card catalogs have provided for libraries. It would assure that every computer file would have accompanying it the basic minimum of documentation provided in a uniform manner. It would differ from the card catalog in that the information provided would be substantially greater and would be an integral part of the computer file itself.

Although such documentation may be difficult to develop for existing data, it might well be put into practice for data which is being generated currently by different government agencies, and which will serve as the basis for future archives. Perhaps this is the sort of matter to which the Office of Statistical Standards might give consideration in their approval of funds for statistical activities. Agencies requesting funds for statistical activities could be required to put their basic files in order and to document them in a systematic manner. Even on existing collections of machine readable data, it might in the long run be advantageous if all known documentation were incorporated in a systematic manner in a documentation header added to the existing computer file.

The problem of privacy

Preserving basic micro-data records and making them available for use does raise questions about the possible invasion of privacy. Since this topic will be the object of discussion during the afternoon session, I will not dwell upon it, but I would like to raise one aspect which I feel may not be covered.

Recently I have been impressed that the problem of individual privacy involves much more than the collection of micro-data in a central research facility. The threat to individual privacy through the collection and misuse of information exists irrespective of whether the information is brought together in one central location. Some of the most flagrant abuses that have occurred in recent periods are due to certain government agencies demanding what is essentially improper information. Unfortunately at the present time many government bodies operate with considerable independence and little supervision in this respect. They have freedom to hire investigators to collect information from other government agencies, private business, and private individuals, and to exchange information with other groups. There is small wonder that a sense of uneasiness about privacy exists. The individual, furthermore, does not know what information is held on him, and he does not have the opportunity to correct erroneous or false information. As a nation we like to think that there is such a thing as due process of law, and that before an individual is convicted there will be a proper weighing of the evidence and he will be given a chance to face his accusers. One of the functions of the courts is to bar inadmissible evidence. No such process operates in the case of administrative actions by government agencies which may affect individuals' lives fully as much as the results of many court actions. The greatest harm that is done to individuals is in precisely those agencies and organizations where information collection is unhampered and secretive, and information is used against the individual in terms of decisions which seriously affect him. One should not be under the illusion that perpetuation of the existing highly decentralized system without controls provides any protection for the individual.

There are signs that this situation may be changing. Recently, for example, New York State has set up a centralized identification and intelligence system. Prior to the development of this system, there were over 70 million files in various agencies of criminal justice in New York. These were held by police departments, prosecutors, criminal courts, and probation, correction, and parole agencies, all of whom dealt with individuals who came within the jurisdiction of the law. They include, of course, local agencies as well as those of the state government. In all, some 3600 agencies were involved, of which over 600 were police departments. Under the decentralized system of duplicate files, the cost of maintaining the files was very great, and there was no agreement about the kind of evidence which it was proper to keep in the files. In many cases useful information could not be brought to bear upon a pressing problem. The files often were barren of material they should contain, and instead held a collection of newspaper clippings, loose notes, and unverified and irrelevant information. Violation of files was frequent. Police reporters looking for a good story were often given free access to files on suspects, and as a result were able to publish some very interesting but in many cases misleading, irrelevant, and damaging pieces of information. Those police chiefs who tried to protect the confidentiality of their files received poor press treatment, so that they would be encouraged to cooperate with the press more fully in the future.

With the establishment of the statewide identification and intelligence system, one of the first steps was to define what material should be contained in it. Unreliable and inadmissible evidence was excluded. Each agency contributing information was given the right to specify what other agencies should be allowed access to the information. Each administrative unit in the system has access only to that kind of information in the central file which it has been agreed in advance is proper. The intelligence system, furthermore, keeps a record of all information provided to each individual user so that violations of rules of disclosure will be apparent.

One of the strongest supporters of the new identification and intelligence system is the Civil Liberties Union, which has argued that the new system regularizes the kind of information available, increases its accuracy, and protects the rights of the individuals involved. The identification and intelligence system itself has no operational or administrative responsibilities. It has been set up as an independent and impartial information utility designed to meet legitimate requests for information and to protect against the misuse of information. It is interesting to note in this connection that the new agency does have the responsibility of providing data within proper disclosure rules for legitimate research work in the fields of crime, juvenile delinquency, mental health, and other concerns of social research.

The key to the problem of protecting privacy is not to depend blindly upon the inefficiency which may accompany decentralization - in many instances the decentralization may result in flagrant abuses which are difficult to uncover simply because of the extent of the decentralization. If we are to correct the present abuses which do constitute invasion of individual privacy, the system of decentralized data files should not be allowed to continue. Explicit consideration should be given to precisely what kinds of information different government agencies should be permitted to gather and keep, and positive steps should be taken to see that the confidentiality of information is protected and its misuse prevented. The only systematic way to undertake such a reform is to set up an independent non-operating agency specifically concerned with the task of monitoring the information system and preventing its abuse. One of the first requirements would be to exclude from the files of all government agencies improper information which is now entered. Although the complaint will be made that this would result in the loss of important sensitive material needed for crime detection or security purposes, this same charge has been made in the past with respect to the question of the admissiblity of certain types of evidence in the courts. Thus for example confessions obtained improperly have been ruled inadmissible. In similar manner, casual rumors obtained by interviews with neighbors might also be considered improper for inclusion in an individual file. In any event, every individual should have a right to see the information contained in the files relating to him. The independent non-operating agency should be concerned with enforcing compliance by government bodies with standard codes developed to protect the rights of individual privacy.

At the same time, it would be possible for such an agency to bring together for statistical and research purposes the basic files of the different data-gathering agencies. This does not mean, of course, that anyone in any agency could push the proverbial button and get any information he wished. The Internal Revenue Service still should not be able to get lists of names of individuals who respond to Census enumerators. Similarly, individual tax return information should not be made available to other government agencies. For statistical and research purposes, however, it should be possible to tap the basic information of all agencies in order to further our knowledge about the conditions and operation of our society. Thus for example, although medical records are of the most sensitive nature, studies of cancer in relation to air polution might well find it useful to process medical and demographic data together in the interests of scientific research.

When the topic of a Federal data center was initially discussed, the term "data bank" was widely used by Congress and the press. At that time I strongly opposed this concept, since it suggested huge masses of dormant information lying lifeless - a data morgue. Recently, however, I have changed my view considerably, and feel that the concept of data bank may be appropriate after all. A centralized non-operating data center could function like a bank, in the sense that it would hold the active files of each of the government agencies much in the same way that banks hold the demand deposits of their customers. In the case of the bank, there is extremely active day to day operation. The bank keeps the accounts for each customer, and allows him to withdraw funds equivalent to what he had previously deposited. Normally a bank does not confuse the accounts of its customers unduly, and does not provide one depositor with the funds which belong to another depositor. Basic to the concept of a bank, furthermore, is the use of the deposits for other purposes so long as this does not infringe upon the rights of the original depositor. A data center could be a bank, in the sense that it would receive deposits of information from various agencies and provide this information to them on call. The present development of the computer is in this direction, and this is in fact the mode of operation of the New York State identification and intelligence system. Individual law enforcement agencies do not keep their own files, but through remote consoles they are provided the necessary information from the central computer. The data bank, thus, would become an information utility, not unlike the utilities which now provide telephone service and electricity to the various government agencies. By pooling the central files of all agencies with respect to data considerable economies could be achieved, and confidentiality could be preserved.

Many of the hearings have been concerned with the possible dangers of abuse of a centralized data system through perversion by high government officials, secretive misuse by bureaucrats, outright fraud by those hadnling the information, or theft by those on the outside. Banks have successfully faced all of these problems. To most people money is considerably more attractive than information, and it is more fungible.

Although there have been occasional breaches of bank security, few people decide that mattresses are safer and no one seriously suggests that bringing money together increases risk to such a degree that we should not have banks. The U.S. government has shown little sign of being corrupt in financial matters. Few presidents have - at least in the United States - been caught dipping their hands in the till of the Treasury. Few Treasury Secretaries or Federal Reserve Board Chairman have been caught taking funds. Even those handling the funds more directly in the Treasury and Federal Reserve banks are generally considered above suspicion. It is possible that a government employee at some stage or other may engage in outright theft, but the incidence of this, or of armed robbery, is quite low. Our gold in Fort Knox is drained only legally, with full knowledge of the national and international community. It is absurd to think that the task of providing security for detailed information is more difficult than that of providing it for money. I cannot imagine Bonnie and Clyde holding up a computer center for data.

Seriously, there is need for a general service statistical agency which can serve as an information utility for the federal agencies and for research purposes and at the same time provide the vehicle for protecting individual rights to privacy. Out of such a system it would be possible to create well documented sets of information needed for many purposes. The question of when data becomes archival is not easy to answer. Perhaps the best answer is that this occurs as soon as the data are created, and that the question of how to preserve data should be considered at the time of its creation, not later.

A. The Precision Instrument Company, of Palo Alto, California, has developed a "Unicon" Laser Mass Memory Recording and Reproduction System. This system, now on the market, provides a means for reliably, permanently and economically recording and reproducing digital data. The Unit employs a laser to vaporize minute holes 4 microns wide by 3 microns long in a metallic surface, sandwiched between two 31" x 4.75" Mylar strips used as the recording medium. In this manner 2.9 billion bits of digital information is recorded in parallel data tracks, with about 11,200 tracks on a recording strip. The strip is mounted on a cylindrical surface (dual drum) for recording or reproducing. The maximum transfer rate is approximately four million bits per second. Between 15 to 20 reels of magnetic tape or 400,000 sheets of 8½ x 11 paper can be stored on one recording strip. The Unit includes a programmable recorder control subsystem which can provide a hardware and software interface compatible with a specified computer system. Since the Mylar strip is a permanent storage medium, while magnetic tape is not, records managers are very interested in the possibilities of this device. One disadvantage is cost. With a "slave" computer the total unit costs about $750,000.

B. The Sylvania Digital Instrumentation Recorder LF 500 is a multitrack laser optical film recorder system that can compact digital signals. It is designed to record high volumes of data, 35 to 40 reels of magnetic tape onto one reel. It can record in a single inch of film over 180,000 bits of information. Each film contains 36 tracks on a 8 mm strip of microfilm 2000' long, wound onto a 10½" reel. It records at a rate of ten million-bits/second. The recorder is similar to a magnetic tape recorder, with the exception that after recording the film must be processed (developed) by conventional photographic means and would take approximately one half hour per reel. The result is a fairly permanent record lasting forty to fifty years. The film is unaffected by atmospheric or electromagnetic conditions if stored under environmental controls. Magnetic tape can be reconstructed from the signals recorded on the microfilm. Cost per unit is about $65,000, with a six month lead time on delivery.

C. The Magic of Holography

Holography received its start in 1947 when Dr. Dennis Gabor of the Imperial College of Science and Technology in London, tried to improve the resolution of the electron microscope. Dr. Gabor set forth the principles of wavefront reconstruction and coined the name "Holography" (in Greek it means the total information). His major problem was an inadequate light source. In 1960, the needed breakthrough came with the development of a highly coherent light source, the monochromatic laser, i.e., a single "pure" color (such as red) light source. By using this light source, 3 dimensional holographic images can be formed by splitting the laser beam. One part of the beam focuses on the subject and is reflected to the photographic plate. The second or reference beam is reflected by a special mirror at right angles and is also reflected onto a photographic plate. Both beams join (interact coherently) to form a pattern on the photographic plate. When the developed plate is illuminated by laser light, one sees a 3D image of the original subject seemingly hanging in space. Using the techniques of laser holography may give rise to a potentially more significant technology; that of dense holographic data storage equipments. Scientists at Bell Telephone, IBM, Xerox, and RCA laboratories have developed read-only memories on tiny glass cubes .062 inches cube. Each cube can hold up to 10,000 pages. These cubes are made up into a 6 inch plate and can store up to 1 trillion bits of data. It has been proposed by Bell Telephone Labs that a read-write-erase holographic memory be developed that would have fast random access, large capacity and would have a page-organized memory. It is expected that this technique would make possible holographic laser mass memories in the order of trillions of bits of data per array (6" x 6") in about eight years.